Understanding Logistic Regression: A Step-by-Step Guide Using Stata

Authored by

Mohammed looti

November 8, 2025

Logistic Regression is a foundational statistical technique specifically employed for modeling the relationship between a set of independent variables and a categorical or binary response variable. Unlike traditional linear regression, which forecasts a continuous numeric outcome, logistic regression is designed to estimate the probability that a specific event will occur. This is achieved by transforming the outcome probability using the logistic function, ensuring the resulting prediction always falls within the valid probability range of 0 to 1. This method is crucial when analyzing outcomes that possess only two potential states, such as pass/fail, disease presence/absence, or purchase/no purchase.

A clear understanding of the prerequisites for deploying this modeling approach is essential for accurate scientific inference. Whenever a research question centers on predicting the likelihood of a two-category event, logistic regression offers the most statistically appropriate framework. The model relies on the mathematical properties of the sigmoid function to map complex linear combinations of input variables onto a quantifiable probability value, thereby providing interpretable metrics of risk and association.

The Core Utility of Logistic Regression in Applied Research

The primary strength of Logistic Regression lies in its ability to predict probabilities associated with dichotomous outcomes across a vast array of disciplines, spanning public health, financial risk assessment, and behavioral social sciences. By quantifying the influence of predictor variables on the odds of the outcome occurring, researchers gain valuable, actionable insights into complex phenomena. For instance, this model is routinely used to identify factors that increase the risk of specific medical conditions or to predict phenomena like corporate default or consumer attrition (customer churn) in business analytics.

It is imperative to apply logistic regression when the response variable is inherently binary. Consider the following common scenarios where this statistical necessity dictates the choice of model:

In preventative medicine, we might investigate how specific lifestyle factors--such as daily caloric intake, intensity of exercise, and body mass index--influence the probability of experiencing a major cardiac event. The dependent outcome variable, **heart attack occurrence**, is fundamentally binary: it either happens (1) or it does not happen (0).

Within the field of education, analysts frequently assess how detailed academic metrics, including cumulative grade point averages, scores on standardized entrance exams, and the completion rate of advanced placement courses, impact the probability of gaining acceptance to a highly selective university program. The response variable, **university acceptance**, is strictly binary: accepted (1) or rejected (0).

For computational linguistics and digital security analysis, a model might be developed to

determine whether email features, such as total character count, the inclusion of certain keywords, or characteristics of the subject line, affect the probability that an email is accurately classified as spam. Here, the response variable, **spam classification**, yields only two possible outcomes: spam or legitimate.

This comprehensive guide is meticulously structured to walk you through the exact procedures required to execute and rigorously interpret a logistic regression model utilizing the robust statistical software, **Stata**. We will employ a classic public health dataset to clearly illustrate the entire analytic process, from initial data loading to the final, professional reporting of results.

Case Study: Analyzing Maternal Factors and Low Birthweight Risk

To provide a tangible demonstration of this powerful methodology, we will tackle a critical public health research question: assessing whether a mother's age and her specific smoking habits during pregnancy significantly alter the probability of delivering a baby classified as having a low birthweight. Medically, low birthweight is defined as a weight below 2,500 grams (approximately 5 pounds, 8 ounces) and is strongly correlated with increased risks of infant mortality and long-term developmental challenges.

In this specific investigation, the outcome of interest--low birthweight--is coded as a binary variable (yes=1/no=0). Consequently, performing a logistic regression analysis represents the only statistically sound choice for inference. We formulate the hypothesis that both advanced maternal age and active smoking status will serve as statistically significant explanatory variables influencing the likelihood of this adverse outcome. The goal of the study is twofold: to determine the nature of the relationship and to quantify the magnitude of the associated risk using the widely accepted metric of **Odds Ratios**.

We will systematically follow the standard analytical steps within Stata to analyze the predefined dataset known as *lbw*, which compiles detailed information pertaining to 189 distinct maternal cases. Adherence to these steps guarantees a transparent, reliable, and easily replicable analysis.

Step 1: Data Acquisition and Initial Loading in Stata

The foundational step in any statistical endeavor is the efficient acquisition and loading of the necessary data into the chosen analytical environment. For the purpose of this tutorial, the *lbw* dataset is conveniently hosted and made available directly through Stata's official public data repositories. Executing this step correctly ensures that all subsequent commands operate on the intended, validated data structure.

To load this dataset instantly into your current Stata session, simply navigate to the Command window and precisely enter the following instruction:

use http://www.stata-press.com/data/r13/lbw

Once this command has been successfully executed, the dataset, which comprises variables detailing maternal characteristics and birth outcomes, will be actively loaded into memory, making it immediately available for detailed exploration and rigorous modeling procedures. This standard practice ensures ease of accessibility and reproducibility for all users.

Step 2: Generating Descriptive Statistics and Identifying Variables

Prior to initiating the formal regression model, it is mandatory best practice to generate a comprehensive summary of the underlying data. This preliminary descriptive analysis provides immediate insight into variable distributions, observation counts, and the range (minimum and maximum) of values. Such exploration is crucial for identifying any potential data quality issues, such as erroneous outliers or unexpected variable ranges, thereby confirming the integrity of the variables designated for use in the regression.

To obtain a rapid, initial overview of all variables contained within the loaded *lbw* dataset, execute the following straightforward command:

summarize

The output generated by the `summarize` command confirms that the dataset contains a total of 11 variables measured across 189 distinct observations. Although the dataset is comprehensive, our specific logistic regression model will be strictly focused on the three variables determined to be critical for answering our research question: the dependent outcome variable and the two primary predictor variables.

```
. use http://www.stata-press.com/data/r13/lbw
(Hosmer & Lemeshow data)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	189	121.0794	63.30363	4	226
low	189	.3121693	.4646093	0	1
age	189	23.2381	5.298678	14	45
lwt	189	129.8201	30.57515	80	250
race	189	1.846561	.9183422	1	3
smoke	189	.3915344	.4893898	0	1
ptl	189	.1957672	.4933419	0	3
ht	189	.0634921	.2444936	0	1
ui	189	.1481481	.3561903	0	1
ftv	189	.7936508	1.059286	0	6
bwt	189	2944.286	729.016	709	4990

The variables relevant to our upcoming analysis must be precisely defined, particularly concerning their numerical coding structures, as the interpretability of logistic regression results hinges on the accurate understanding of these inputs:

low - This serves as our binary response variable, indicating whether the baby experienced low birthweight. It is coded such that 1 signifies the occurrence of low birthweight (the positive outcome) and 0 signifies the non-occurrence (normal birthweight).
age - This is the continuous explanatory variable, measured as the mother's age in years at the time of delivery.
smoke - This is a binary explanatory variable indicating whether the mother engaged in smoking during the pregnancy. It is coded as 1 if the mother smoked and 0 if she did not smoke.

Step 3: Executing the Logistic Regression Model in Stata

With the dataset loaded and the key variables clearly identified and confirmed, we are now ready to execute the core logistic regression model. In the Stata environment, the dedicated command for this procedure is `logit`. The necessary syntax requires the user to specify the binary dependent variable first, immediately followed by the complete list of independent (predictor) variables. The `logit` command then employs the method of maximum likelihood estimation to calculate the coefficients, which quantify the change in the log-odds of the outcome associated with a one-unit increase in each predictor.

To construct the model predicting the probability of low birthweight (low) based on the mother's

age (age) and her smoking status (smoke), execute the following command:

logit low age smoke

Upon successful execution, Stata will generate an exhaustive table containing critical information regarding the overall model fit and the specific coefficients for every predictor included. This detailed output forms the absolute basis for statistically interpreting the relationships between our maternal factors and the likelihood of a low birthweight delivery. Thoroughly understanding the various metrics presented is paramount for drawing statistically valid and meaningful conclusions.

```
. logit low age smoke

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -113.66733
Iteration 2:   log likelihood = -113.63815
Iteration 3:   log likelihood = -113.63815

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =       7.40
                                                Prob > chi2       =     0.0248
Log likelihood = -113.63815                     Pseudo R2         =     0.0315
```

low	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0497792	.031972	-1.56	0.119	-.1124431	.0128846
smoke	.6918486	.3218061	2.15	0.032	.0611202	1.322577
_cons	.0609051	.7573199	0.08	0.936	-1.423415	1.545225

Step 4: Interpreting Key Coefficients and Assessing Significance

The primary Stata output delivers several crucial pieces of information necessary for statistical inference, namely the raw coefficients (Coef.), their corresponding standard errors, the calculated test statistic (z), and the associated p-value (P>|z|). Interpreting these statistics allows us to evaluate both the direction and the statistical significance of each predictor variable's effect. However, for practical and intuitive interpretation, the **Odds Ratio** is often preferred, as it is calculated by simply exponentiating the coefficient (exp(Coef.)).

We must analyze the findings for the continuous variable (age) and the categorical variable (smoking status) separately, paying careful attention to their respective impacts on the odds of low birthweight:

Coefficient for Age (Coef): The coefficient is reported as -.0497792. This negative value indicates

the magnitude of change in the log-odds of having a low birthweight baby for every single one-year increase in maternal age, provided the smoking status remains constant. The inverse relationship suggests that older mothers, all other variables being equal, experience slightly reduced odds of the outcome. Translating this log-odds change into an intuitive Odds Ratio yields: exp(-.0497792) = 0.951. An Odds Ratio below 1.0 signifies a reduction in the outcome odds. Specifically, for each additional year of age, the odds of delivering a low birthweight baby are estimated to decrease by approximately 4.9%.

P-value for Age (P>|z|): The computed p-value is 0.119. This value represents the probability of observing a Z-statistic as extreme as -1.56 under the null hypothesis (i.e., if age truly had no effect on the outcome). Because 0.119 exceeds the conventional significance threshold (alpha = 0.05), we are compelled to conclude that maternal age is **not** a statistically significant predictor of low birthweight risk within this specific sample.

The interpretation for the categorical smoking variable requires a direct comparison between the two coded groups (smokers vs. non-smokers):

Coefficient for Smoking (Coef): The coefficient is reported as 0.6918486. This positive coefficient strongly indicates that smoking significantly increases the log-odds of the adverse outcome. The derived Odds Ratio is calculated as exp(0.6918486) = 1.997. This powerful metric implies that, when holding maternal age constant, a mother who smokes during pregnancy has nearly double (1.997 times) the odds of having a low birthweight baby compared to a mother who refrains from smoking during pregnancy. This constitutes a substantial and clinically relevant increase in risk.

P-value for Smoking (P>|z|): The associated p-value is 0.032. Since 0.032 falls below the standard significance level of 0.05, we confidently determine that smoking status is a **statistically significant** predictor of low birthweight. This strongly suggests that the observed elevated risk is robust and unlikely to be the result of random sampling variation.

Step 5: Formal Communication and Reporting of Final Results

The concluding stage of the statistical process requires synthesizing the analytical findings into a clear, concise, and formally structured summary suitable for dissemination in academic journals, grant applications, or professional reports. Standard statistical reporting necessitates the inclusion of the analysis type performed, the sample size used, and precise results concerning the statistical significance and estimated effect size (Odds Ratio) for every predictor variable.

When reporting the outcomes derived from Stata or similar software, it is mandatory practice to include both the test statistic (z) and the exact p-value for each predictor. This provides the necessary evidential foundation to support the claims of significance or non-significance, allowing the reader to independently verify the model's conclusions.

The following is a standard example of a formal narrative write-up summarizing the key findings

derived from the logistic regression analysis we performed:

A logistic regression was performed to evaluate the effects of maternal age and smoking habits during pregnancy on the probability of delivering a baby with a low birthweight. A total sample of 189 mothers was included in this analysis.

Results demonstrated conclusively that maternal smoking status was a statistically significant predictor of low birthweight (z = 2.15, p = .032). Mothers who reported smoking had nearly double the odds of having a low birthweight baby compared to non-smokers, while controlling for the mother's age (OR = 1.997, 95% CI). In contrast, maternal age was not found to have a statistically significant relationship with the probability of low birthweight (z = -1.56, p = .119). Although the coefficient suggested a marginal decrease in odds with increasing age (OR = 0.951), this measured effect was not statistically distinguishable from zero.

This structured reporting style effectively communicates the critical relationship between the predictors and the binary outcome, providing both the easily interpretable magnitude of the effect (Odds Ratio) and the robust statistical evidence necessary for formal inference (z and p-values).